Biased Reduced Sampling: Detectability of an Attribute and Estimation of Prevalence

Todd Graves and Michael Hamada

Statistical Sciences

Los Alamos National Laboratory

Los Alamos, NM 87545

09.30.03 2200

Abstract

In surveilling a population, detection of systems with an attribute of interest and estimation of the prevalence of the attribute in the population are two main goals. Cost constraints may severely limit the fraction of systems whose components can be sampled and tested. Biasing the sampling to increase the probability of choosing a component with an attribute of interest ameliorates the impact of reduced sampling. In this paper, we consider the impact of biased reduced sampling on detection and propose an approach for estimating the prevalence of the attribute in the population which properly accounts for the biasing. The proposed method is illustrated with a simulated example.

Key Words: Bayesian methods, extended-hypergeometric and hypergeometric distributions, MCMC.

Introduction

Populations are surveilled to make sure they are healthy. For example, missile and weapon stockpiles are surveilled to make sure that they work when needed. In surveilling a population, there are two goals: (1) detection – can you find a system which has an attribute of interest? and (2) estimation – if you find such a system, how many other systems in the population have this attribute?

Surveillance can take the form of periodically sampling a number of systems from the population in which the sampling is completely random, i.e., each system in the population has the same probability of being chosen. For multiple component systems, it may be too expensive to inspect every component. Hence, for some components, only a subset of them can be tested and these are sampled from the systems chosen in the original sample. To support the detection goal, the reduced sampling can be ameliorated by biasing the sampling to find components which have an attribute of interest. We refer to this as biased reduced sampling. The sampling is still stochastic but components with the attribute have a higher probability of being selected.

In this paper, first we consider the impact of biased reduced sampling on detection. Next, we consider estimation of prevalence using such data for which the biasing needs to be accounted. A Bayesian approach is proposed and then illustrated with an example. The paper concludes with a discussion.

Impact on Detectability of an Attribute

Consider a large population in which a portion p of the systems have an attribute. Let n_1 and n_2 be the first and second stage sample sizes, respectively. The first stage sample is completely random, i.e., unbiased. The second stage sample is stochastic but systems with an attribute have a higher chance of being selected. Suppose that we have m such samples in which y_i, the number of systems in the ith second stage sample with the attribute, are observed.

The following model is assumed for the first and second stage sampling:

$$K_i \sim Binomial(n_1, p), \tag{1}$$

where K_i is the unknown number of systems out of n_1 having the attribute in the i'th first stage sample;

$$y_i \sim Extended - hypergeometric(K_i, n_1 - K_i, n_2, \theta), \tag{2}$$

where y_i is the observed number of systems out of n_2 having the attribute in the second stage sample. When $\theta = 1$, the extended-hypergeometric distribution reduces to the hypergeometric distribution which holds for complete random sampling. When $\theta > 1$, systems with the attribute are favored in the sampling. The probability mass function has the following form:

$$P(y_i = y) = \frac{\left(\begin{array}{c} n_2 \\ y \end{array} \right) \left(\begin{array}{c} n_1 - n_2 \\ K_i - y \end{array} \right) \theta^y}{\sum_{j=\max(0, n_2 - n_1 + K_i)}^{\min(n_2, K_i)} \left(\begin{array}{c} n_2 \\ j \end{array} \right) \left(\begin{array}{c} n_1 - n_2 \\ K_i - j \end{array} \right) \theta^j}, \tag{3}$$

for $y = \max(0, n_2 - n_1 + K_i), \ldots, \min(n_2, K_i)$. See Table 1 which demonstrates

that when $\theta > 1$, the probabilities for sampling a system with an attribute is higher than for completely random sampling.

As an alternative, the noncentral-hypergeometric distribution arises when systems with the attribute have weight ω, systems without the attribute have weight 1 and systems are chosen with probabilities proportional to their weights. The noncentral-hypergeometric probability mass function has the form:

$$P(y_i = y) = \begin{pmatrix} K_i \\ y \end{pmatrix} \begin{pmatrix} n_1 - K_i \\ n_2 - y \end{pmatrix} \int_0^1 (1 - z^{\omega\gamma})^y (1 - z^{\gamma})^{n_2 - y} dz, \qquad (4)$$

where $\gamma = 1/(\omega(K_i - y) + \{(n_1 - K_i) - (n_2 - y)\})$. While the latter distribution may be more interpretable, both distributions are similar when the first stage sample contains equal number of systems with and without the attribute. Note in Table 2 that the noncentral-hypergeometric probabilities are similar when $\omega = \theta$ for $n_1 = 10$, $K = 5$ and $n_2 = 3$ to the extended-hypergeometric probabilities in Table 1. In making inferences with such data, we will see that the extended-hypergeometric distribution is preferable because its probability mass function involves a sum rather than an integral. The extended and noncentral hypergeometric distributions are both discussed in Johnson and Kotz (1969).

In surveilling a population, one rationale for determining a sample size is as follows. If the proportion p is 0.10, what sample size is needed so that in two years there is a 0.90 probability of randomly sampling a system with the attribute? The corresponding sample size is 22 systems or 11 per year. To

Table 1. Extended-hypergeometric Probabilities for $n_1 = 10$, $K = 5$, $n_2 = 3$ ($\theta = 1$ is hypergeometric)

	θ				
y	0.1	0.5	1.0	1.5	2.0
0	0.645	0.205	0.083	0.043	0.026
1	0.322	0.513	0.417	0.324	0.256
2	0.032	0.256	0.417	0.487	0.513
3	0.001	0.026	0.083	0.146	0.205

Table 2. Noncentral-hypergeometric Probabilities for $n_1 = 10$, $K = 5$, $n_2 = 3$ ($\omega = 1$ is hypergeometric)

	ω				
y	0.1	0.5	1.0	1.5	2.0
0	0.684	0.223	0.083	0.039	0.020
1	0.290	0.522	0.417	0.310	0.246
2	0.025	0.232	0.417	0.492	0.514
3	0.0003	0.023	0.083	0.160	0.220

see the impact of reduced biased sampling when $n_1 = 11$, see Table 3 which presents the probability of sampling a system with an attribute when the number of systems with attributes in the first stage sample is $K = 1, \ldots, 10$ and a second stage sample of size one ($n_2 = 1$) is taken.

To see the impact of reduced biased sampling on detection, we need to evaluate the probability that a second stage sample contains at least one system with an attribute for different proportions p. This probability can be expressed as

$$P(y_i \geq 1) = \sum_{j=1}^{n_1} P(y_i \geq 1 | K_i = j, n_1, n_2, \theta) P(K_i = j | p, n_1), \qquad (5)$$

where the first term is the sum of extended-hypergeometric probabilities and

Table 3. Probability of sampling a system with an attribute in the second stage when n_2=1 for first stage where n_1=11 and K is the number of systems with an attribute ($\theta = 1$ is hypergeometric)

	θ						
K	0.1	0.5	1.0	1.5	2.0	5.0	10.0
1	0.010	0.048	0.091	0.130	0.167	0.333	0.500
2	0.022	0.100	0.182	0.250	0.308	0.526	0.690
3	0.036	0.158	0.273	0.360	0.429	0.652	0.789
4	0.054	0.222	0.364	0.462	0.533	0.741	0.851
5	0.077	0.294	0.455	0.556	0.625	0.806	0.893
6	0.107	0.375	0.545	0.643	0.706	0.857	0.923
7	0.149	0.467	0.636	0.724	0.778	0.897	0.946
8	0.211	0.571	0.727	0.800	0.842	0.930	0.964
9	0.310	0.692	0.818	0.871	0.900	0.957	0.978
10	0.50	0.833	0.909	0.937	0.952	0.980	0.990

the second term is a binomial probability. See Table 4 which provides these probabilities for n_1=11 and $n_2 = 1, 4, 7$ for various proportions p and biasing determined by θ. The first stage probabilities are given in parentheses. Table 4 shows that biasing the second stage samples helps to ameliorate the reduced sampling; for example, when $\theta = 5$ and $n_2 = 7$ the detectability is no more than 0.04 less than if $n_2 = 11$.

Estimation of Prevalence

In this section we consider estimation of p using biased reduced sampling data. We assume the statistical model given in the previous section for m second stage samples yielding the data $y_1, \ldots, y_m$. Without some completely

randomly sampled data, there is no information about the amount of biasing θ. Hence, we must perform the estimation with a specified value of $\theta > 1$. To do the estimation, we propose using a Bayesian approach.

Bayesian inference provides uncertainty about the unknowns $\boldsymbol{\eta} = (p, K_1, \ldots, K_m)$ through their joint posterior distribution. It does so by combining prior information about $\boldsymbol{\eta}$ with the information about $\boldsymbol{\eta}$ contained in the data. The prior information is described by a probability density $\pi(\boldsymbol{\eta})$ known as the prior density and the information provided by the data is captured by the data sampling model $f(\mathbf{y}|\boldsymbol{\eta})$ known as the likelihood. The combined information is then described by another probability density $\pi(\boldsymbol{\eta}|\mathbf{y})$ called the posterior density. Bayes Theorem provides the way to calculate the posterior density, namely,

$$\pi(\boldsymbol{\eta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\eta})\pi(\boldsymbol{\eta}). \tag{6}$$

For this problem, we only need to specify a prior distribution for p since the K_i are specified by (1) and the likelihood is given by (2). We use the following prior:

$$p \sim Beta(a_0, b_0), \tag{7}$$

for specified a_0 and b_0.

Because there are $m+1$ parameters, we employ an appropriate MCMC (Markov Chain Monte Carlo) method to sample the joint posterior distribution (Gelman et al., 1995) from which inference about the unknown parameter of interest p can be made. For example, the Metropolis-Hastings algorithm (Chib and Greenberg (1995)) combined with Gibbs sampling (Casella and George (1992)) provide a general way to sample from the joint posterior distribution.

If some of the second stage samples are completely random, then a comparison between the completely random and biased random samples provides information about θ. For completely random samples,

$$y_{ri} \sim Hypergeometric(K_i, n_1 - K_i, n_2). \tag{8}$$

Also, we need to specify a prior for θ such as

$$\theta \sim 1 + Lognormal(0, t_0), \tag{9}$$

which is defined on $(1, \infty)$. This assumes that in the second stage, systems with an attribute have a higher probability of being chosen. To be more conservative and also allow for the possibility of systems with an attribute having a lower probability of being chosen, the prior

$$\theta \sim Lognormal(0, t_0) \tag{10}$$

can be employed. Then we apply Bayes Theorem as before using the appropriate MCMC method to provide inference about p and θ.

An Example

To illustrate the estimation of prevalence, first consider the case when θ is known. Suppose that $\theta = 4$. Table 5 presents simulated data for $m = 15$ samples when $p = 0.20$. Note that the K_i (number of systems in the first stage sample with an attribute) are unknown and only the y_i (number of systems in the second stage sample with an attribute) are observed.

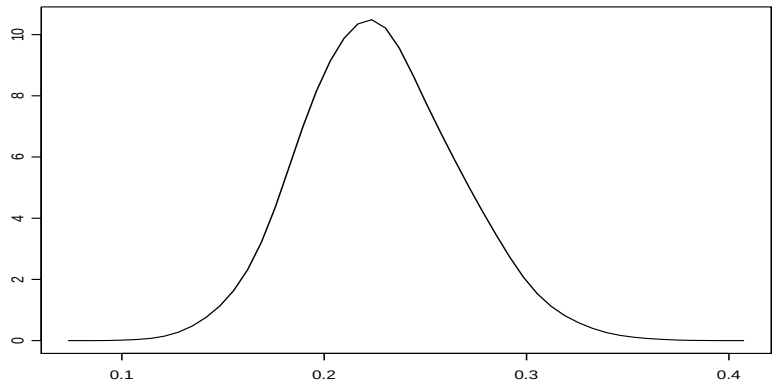

Figure 1: Posterior for p for known $\theta = 4$.

The Bayesian analysis described in the previous section was performed on these data using a $Beta(0.5, 0.5)$ prior for p; i.e., a non-informative prior for p. We implemented the MCMC algorithm using the YADAS statistical modeling environment (Graves, 2001, 2003a,b). WinBUGS (Spiegelhalter, Thomas, and Best (2000)) was not used because it cannot handle the extended-hypergeometric distribution. See the posterior obtained for p in Figure 1. The posterior 0.05, 0.50, 0.95 quantiles for p are 0.183, 0.226, 0.291, respectively. Thus, an estimate for p using the median of the posterior is 0.226 compared with the true p of 0.20. Estimates of the K_i are also obtained in case they are of interest.

Consider what would have happened if the biased sampling in the second stage had been ignored. There are 25 out of 60 systems with the attribute in the second stage samples which yields an estimated p of $25/60 = 0.417$, an overestimate.

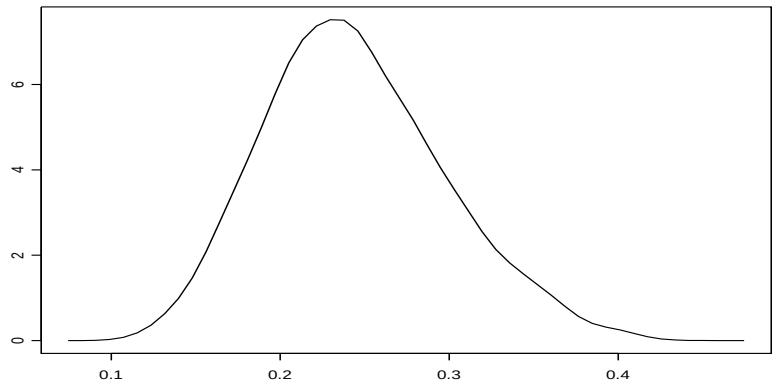

Figure 2: Posterior for p when θ is unknown.

Now consider the situation where θ is not known, but some second stage samples are available which are completely random. See Table 6 which presents 10 completely random second stage samples. The comparison between the biased and unbiased second stage samples allows θ to be estimated. A Bayesian analysis was performed using the following priors:

$$p \sim Beta(0.5, 0.5) \text{ and } \theta \sim Lognormal(0, 1).$$

See the resulting posteriors for p and θ in Figures 2 and 3. The posterior 0.05, 0.50, 0.95 quantiles for p are 0.178, 0.240, 0.340, respectively. Thus an estimate for p using the median of its posterior is 0.240 as compared with the true p of 0.20. The posterior 0.05, 0.50, 0.95 quantiles for θ are 1.508, 3.337, 7.435, respectively, as compared with the true θ of 4.

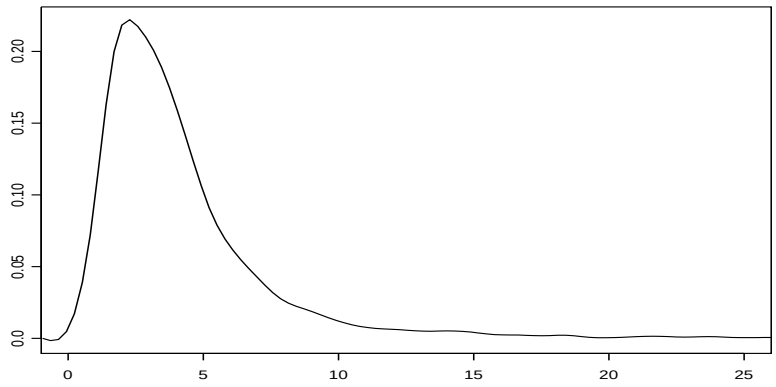

Figure 3: Posterior for θ.

Discussion

In this paper, we have shown how reduced sampling can be ameliorated by biasing; i.e., systems with an attribute of interest have an increased probability of being sampled. The biasing if ignored causes problems in estimating the prevalence of the attribute in the population. We have shown how a Bayesian approach easily accounts for the biasing when either the extent of the biasing is known or if unknown unbiased data are also available which allow the extent of the biasing to be estimated.

In the example, we assumed that all the first stage samples were the same size as well as second stage samples. This need not be and yet the proposed Bayesian approach handles such unequal sample size situations. In this paper, we assumed that the surveilled population is large so that the number of systems with an attribute in the first stage samples approximately follows a binomial distribution. A topic of future research is to consider

a small population in which the first stage completely random sample is modeled by a hypergeometric distribution.

Acknowledgements

We thank Dee Won for her encouragement of this work.

References

Casella, G. and George, E. (1992). "Explaining the Gibbs Sampler." *The American Statistician* 46, 167–174.

Chib, S. and Greenberg, E. (1995). "Understanding the Metropolis-Hastings Algorithm." *The American Statistician* 49, 327–335.

Gelman, A.B., Carlin, J.S., Stern, H.S., and Rubin, D.B. (1995) *Bayesian Data Analysis*. Boca Raton: Chapman and Hall/CRC.

Graves, T.L. (2001) "YADAS: An Object-Oriented Framework for Data Analysis Using Markov Chain Monte Carlo," Los Alamos National Laboratory Technical Report LA-UR-01-4804.

Graves, T.L. (2003a) "A Framework for Expressing and Estimating Arbitrary Statistical Models Using Markov Chain Monte Carlo," Los Alamos National Laboratory Technical Report LA-UR-03-5934.

Graves, T.L. (2003b) "An Introduction to YADAS," `yadas.lanl.gov`.

Johnson, N.L. and Kotz, S. (1969). *Discrete Distributions.* Boston: Houghton Mifflin.

Spiegelhalter, D.,Thomas, A. and Best, N. (2000), *WinBUGS Version 1.3 User Manual*.

Table 4. Reduced Biased Sampling Plan Properties
(Table entry is probability that at least one system in second stage sample has an attribute. Probability that at least one system in first stage has an attribute when $n_1 = 11$ is given in parentheses.)

	p=0.05 (0.431)						
	θ						
n_2	0.1	0.5	1.0	1.5	2.0	5.0	10.0
1	0.006	0.027	0.050	0.070	0.088	0.165	0.237
4	0.031	0.119	0.185	0.228	0.258	0.340	0.380
7	0.087	0.234	0.302	0.335	0.354	0.397	0.413

	p=0.10 (0.686)						
	θ						
n_2	0.1	0.5	1.0	1.5	2.0	5.0	10.0
1	0.012	0.055	0.100	0.138	0.170	0.300	0.411
4	0.067	0.233	0.344	0.411	0.456	0.571	0.624
7	0.182	0.425	0.522	0.566	0.592	0.645	0.665

	p=0.15 (0.833)						
	θ						
n_2	0.1	0.5	1.0	1.5	2.0	5.0	10.0
1	0.019	0.085	0.150	0.202	0.246	0.411	0.540
4	0.106	0.339	0.478	0.556	0.606	0.726	0.777
7	0.281	0.577	0.679	0.723	0.748	0.797	0.814

	p=0.20 (0.914)						
	θ						
n_2	0.1	0.5	1.0	1.5	2.0	5.0	10.0
1	0.027	0.116	0.200	0.265	0.317	0.502	0.636
4	0.151	0.439	0.590	0.670	0.719	0.829	0.871
7	0.375	0.695	0.789	0.827	0.848	0.887	0.901

Table 5. First and second stage sample data when $m = 15$, $p = 0.20$, $\theta = 4$

Sample	n_1	K	n_2	y
1	11	3	4	2
2	11	3	4	2
3	11	4	4	2
4	11	1	4	1
5	11	4	4	2
6	11	0	4	0
7	11	3	4	1
8	11	1	4	0
9	11	5	4	4
10	11	2	4	2
11	11	3	4	2
12	11	2	4	2
13	11	2	4	2
14	11	2	4	0
15	11	4	4	3

Table 6. First and second stage sample data when $m = 10$, $p = 0.20$, $\theta = 0$

Sample	n_1	K	n_2	y
1	11	1	4	0
2	11	0	4	0
3	11	2	4	1
4	11	2	4	0
5	11	7	4	3
6	11	1	4	1
7	11	4	4	2
8	11	2	4	0
9	11	1	4	0
10	11	3	4	0